EPiC
EPISTEMIC INTELLIGENCE & COMPUTATION LAB

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | College of Computing and Data Science

When One Probability Is Not Enough: On Second-order Uncertainty Formalisms in AI

Siu Lun (Alan) Chau

Epistemic Intelligence & Computation Lab
College of Computing and Data Science, Nanyang Technological University

CISPA seminar

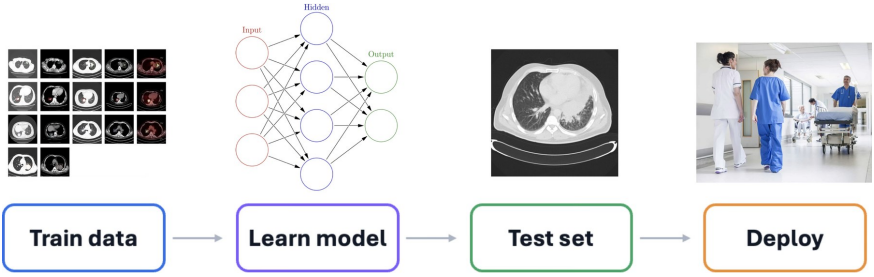
Talk structure

1. Why a single predictive probability is not enough.
2. How second-order formalisms enter.
3. Why a set of probabilities can be read directly as a credal set.
4. How to make ensemble disagreement more meaningful.
5. Which second-order formalism is better?
6. What a controlled comparative study tells us.

Section 1

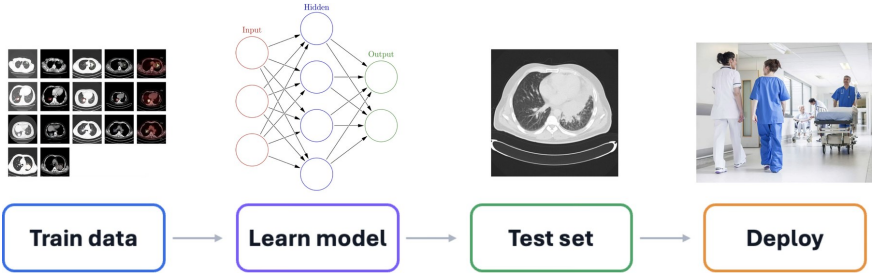
Why first-order prediction is not enough

From prediction to deployment



- Supervised learning predicts an outcome $Y \in \mathcal{Y}$ from an input $X \in \mathcal{X}$, using data $\{(X_i, Y_i)\}_{i=1}^n$.

From prediction to deployment



- Supervised learning predicts an outcome $Y \in \mathcal{Y}$ from an input $X \in \mathcal{X}$, using data $\{(X_i, Y_i)\}_{i=1}^n$.
- Suppose the model returns

$$f_\theta(x) = \text{“stage III”}.$$

- The deployment question is not only whether the label is correct, but whether the prediction is reliable enough to act on.

Is probabilistic prediction the answer?

- A probabilistic classifier outputs a distribution rather than a single label:

$$f_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{Y}).$$

Is probabilistic prediction the answer?

- A probabilistic classifier outputs a distribution rather than a single label:

$$f_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{Y}).$$

- For example,

$$f_\theta(x)[\text{“stage III”}] = \mathbf{0.87}.$$

Is probabilistic prediction the answer?

- A probabilistic classifier outputs a distribution rather than a single label:

$$f_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{Y}).$$

- For example,

$$f_\theta(x)[\text{“stage III”}] = \mathbf{0.87}.$$

- This is already a major improvement over a hard label.

Is probabilistic prediction the answer?

- A probabilistic classifier outputs a distribution rather than a single label:

$$f_\theta : \mathcal{X} \to \mathcal{P}(\mathcal{Y}).$$

- For example,

$$f_\theta(x)[\text{“stage III”}] = \mathbf{0.87}.$$

- This is already a major improvement over a hard label.
- But the reported probability is still only a **first-order** statement about Y.

Takeaway. A probability can describe randomness in the outcome. It does not, by itself, describe our ignorance about the probability.

What first-order probabilistic learning is trying to learn

- Write the data-generating process as

$$P_0(X, Y) = P_0(X)P_0(Y \mid X).$$

What first-order probabilistic learning is trying to learn

- Write the data-generating process as

$$P_0(X, Y) = P_0(X)P_0(Y \mid X).$$

- Under logarithmic loss (strictly proper scoring rule) , the **population risk** is

$$R(f) = \mathbb{E}_{(X,Y)\sim P_0}\big[-\log f(Y \mid X)\big].$$

What first-order probabilistic learning is trying to learn

- Write the data-generating process as

$$P_0(X, Y) = P_0(X)P_0(Y \mid X).$$

- Under logarithmic loss (strictly proper scoring rule) , the **population risk** is

$$R(f) = \mathbb{E}_{(X,Y)\sim P_0}\big[-\log f(Y \mid X)\big].$$

- Conditioning on X, this decomposes as

$$R(f) = \mathbb{E}_X\left[H\big(P_0(\cdot \mid X)\big) + D_{\mathrm{KL}}\big(P_0(\cdot \mid X) \,\|\, f(\cdot \mid X)\big)\right].$$

What first-order probabilistic learning is trying to learn

- Write the data-generating process as

$$P_0(X, Y) = P_0(X)P_0(Y \mid X).$$

- Under logarithmic loss (strictly proper scoring rule) , the **population risk** is

$$R(f) = \mathbb{E}_{(X,Y)\sim P_0}\big[-\log f(Y \mid X)\big].$$

- Conditioning on X, this decomposes as

$$R(f) = \mathbb{E}_X\left[H\big(P_0(\cdot \mid X)\big) + D_{\mathrm{KL}}\big(P_0(\cdot \mid X) \,\|\, f(\cdot \mid X)\big)\right].$$

- Hence the **ideal probabilistic predictor** is

$$f^\star(\cdot \mid x) = P_0(\cdot \mid X = x).$$

Why this ideal is not directly available

Ideally, we would recover

$$f^\star(x) = P_0(\cdot \mid X = x) \qquad \text{for every } x \in \mathcal{X}.$$

Finite data

We minimise empirical risk, not population risk:

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Why this ideal is not directly available

Ideally, we would recover

$$f^\star(x) = P_0(\cdot \mid X = x) \qquad \text{for every } x \in \mathcal{X}.$$

Finite data

We minimise empirical risk, not population risk:

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Modelling assumptions

The true conditional may not belong to the chosen hypothesis class:

$$P_0(\cdot \mid X) \notin \mathcal{H}.$$

Why this ideal is not directly available

Ideally, we would recover

$$f^\star(x) = P_0(\cdot \mid X = x) \qquad \text{for every } x \in \mathcal{X}.$$

Finite data

We minimise empirical risk, not population risk:

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Modelling assumptions

The true conditional may not belong to the chosen hypothesis class:

$$P_0(\cdot \mid X) \notin \mathcal{H}.$$

Optimisation

Training may only find an approximate solution:

$$\widehat{f} \approx \arg\min_{f \in \mathcal{H}} \widehat{R}_n(f).$$

Why this ideal is not directly available

Ideally, we would recover

$$f^\star(x) = P_0(\cdot \mid X = x) \qquad \text{for every } x \in \mathcal{X}.$$

Finite data

We minimise empirical risk, not population risk:

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Modelling assumptions

The true conditional may not belong to the chosen hypothesis class:

$$P_0(\cdot \mid X) \notin \mathcal{H}.$$

Optimisation

Training may only find an approximate solution:

$$\widehat{f} \approx \arg\min_{f \in \mathcal{H}} \widehat{R}_n(f).$$

Takeaway. The learned probability is an estimate of the ideal first-order probability. We still need to express uncertainty about this estimate.

A single probability cannot separate aleatoric and epistemic uncertainty

World A: intrinsic randomness

$$P(Y = 1 \mid x) = 0.5.$$

- The mechanism is known.
- The outcome is genuinely variable.
- High **aleatoric** uncertainty, low epistemic uncertainty.

A single probability cannot separate aleatoric and epistemic uncertainty

World A: intrinsic randomness

$$P(Y = 1 \mid x) = 0.5.$$

- The mechanism is known.
- The outcome is genuinely variable.
- High **aleatoric** uncertainty, low epistemic uncertainty.

World B: model ignorance

$$P(Y = 1 \mid x, \theta_1) = 0,$$
$$P(Y = 1 \mid x, \theta_2) = 1,$$

θ_1, θ_2 both plausible.

- Each mechanism is nearly deterministic.
- We do not know which mechanism applies.
- Low aleatoric uncertainty per model, high **epistemic** uncertainty.

A single probability cannot separate aleatoric and epistemic uncertainty

World A: intrinsic randomness

$$P(Y = 1 \mid x) = 0.5.$$

- The mechanism is known.
- The outcome is genuinely variable.
- High **aleatoric** uncertainty, low epistemic uncertainty.

World B: model ignorance

$$P(Y = 1 \mid x, \theta_1) = 0,$$
$$P(Y = 1 \mid x, \theta_2) = 1,$$

θ_1, θ_2 both plausible.

- Each mechanism is nearly deterministic.
- We do not know which mechanism applies.
- Low aleatoric uncertainty per model, high **epistemic** uncertainty.

Both can produce the same first-order prediction: $\widehat{P}(Y = 1 \mid x) = 0.5$.

Calibration is important, but not enough

- A binary probabilistic predictor is calibrated if

$$\mathbb{P}(Y = 1 \mid \widehat{p}(X) = c) = c.$$

Calibration is important, but not enough

- A binary probabilistic predictor is calibrated if
$$\mathbb{P}(Y = 1 \mid \widehat{p}(X) = c) = c.$$

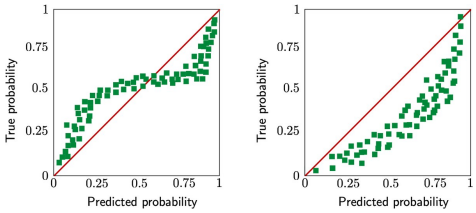
- Calibration asks whether reported probabilities match empirical frequencies.

Calibration is important, but not enough

- A binary probabilistic predictor is calibrated if
$$\mathbb{P}(Y = 1 \mid \widehat{p}(X) = c) = c.$$
- Calibration asks whether reported probabilities match empirical frequencies.
- Good calibration $\neq$ good prediction. Why?

Calibration is important, but not enough

- A binary probabilistic predictor is calibrated if
$$\mathbb{P}(Y = 1 \mid \widehat{p}(X) = c) = c.$$
- Calibration asks whether reported probabilities match empirical frequencies.
- Good calibration $\neq$ good prediction. Why?
- Also calibration is still a **first-order** property of the averaged predictive probabilities. **Fails to tell randomness away from ignorance.**



Reliability diagrams diagnose probability-frequency mismatch; they do not represent uncertainty about the probability itself.

Takeaway. Calibration helps us fix reported probabilities. It does not replace a representation of epistemic uncertainty. It also breaks easily.

Two layers of uncertainty

1. **First-order uncertainty:**

$$P(Y \mid X = x).$$

This describes randomness in Y once the predictive mechanism is known.

Two layers of uncertainty

1. **First-order uncertainty:**

$$P(Y \mid X = x).$$

This describes randomness in Y once the predictive mechanism is known.

2. **Second-order uncertainty:**

uncertainty about $P(Y \mid X = x)$.

This describes incomplete knowledge about the predictive distribution itself.

Two layers of uncertainty

1. **First-order uncertainty:**

$$P(Y \mid X = x).$$

   This describes randomness in Y once the predictive mechanism is known.

2. **Second-order uncertainty:**

$$\text{uncertainty about } P(Y \mid X = x).$$

   This describes incomplete knowledge about the predictive distribution itself.

Takeaway. To distinguish "the world is noisy" from "the model does not know", we need a second-order object.

What should a second-order formalism do?

- Represent **more** than a single predictive distribution.

What should a second-order formalism do?

- Represent **more** than a single predictive distribution.
- Preserve information about which **alternative** predictive distributions remain plausible.

What should a second-order formalism do?

- Represent **more** than a single predictive distribution.
- Preserve information about which **alternative** predictive distributions remain plausible.
- Support **downstream decisions**: rejection, OOD detection, active learning, robust prediction.

What should a second-order formalism do?

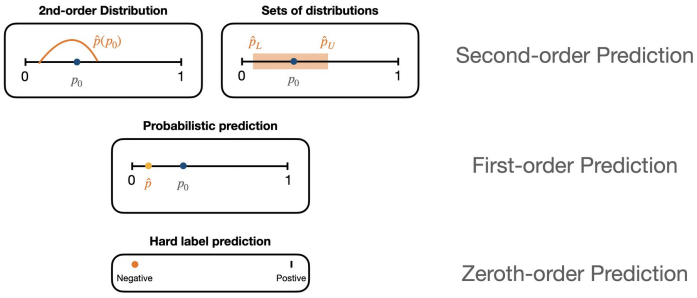
- Represent **more** than a single predictive distribution.
- Preserve information about which **alternative** predictive distributions remain plausible.
- Support **downstream decisions**: rejection, OOD detection, active learning, robust prediction.

The object of interest is no longer just $p(y \mid x)$, but our uncertainty about $p(\cdot \mid x)$.

Section 2

Two ways to represent uncertainty about probabilities

Two common second-order formalisms



Distribution-based

A probability distribution over plausible predictive distributions.

Set-based

A set of plausible predictive distributions, often a credal set.

Distribution-based formalism: Bayesian prediction

- A Bayesian approach places a prior over parameters and updates it using Bayes' theorem:

$$\pi(\theta \mid D) = \frac{p(D \mid \theta)\pi(\theta)}{p(D)}.$$

Distribution-based formalism: Bayesian prediction

- A Bayesian approach places a prior over parameters and updates it using Bayes' theorem:

$$\pi(\theta \mid D) = \frac{p(D \mid \theta)\pi(\theta)}{p(D)}.$$

- The posterior induces a distribution over predictive distributions:

$$\theta \sim \pi(\theta \mid D) \quad \Longrightarrow \quad p_\theta(\cdot \mid x) \in \mathcal{P}(\mathcal{Y}).$$

Distribution-based formalism: Bayesian prediction

- A Bayesian approach places a prior over parameters and updates it using Bayes' theorem:

$$\pi(\theta \mid D) = \frac{p(D \mid \theta)\pi(\theta)}{p(D)}.$$

- The posterior induces a distribution over predictive distributions:

$$\theta \sim \pi(\theta \mid D) \quad \Longrightarrow \quad p_\theta(\cdot \mid x) \in \mathcal{P}(\mathcal{Y}).$$

- Bayesian model averaging returns the first-order prediction

$$\bar{p}(y \mid x, D) = \int p_\theta(y \mid x)\, \pi(d\theta \mid D).$$

Takeaway. The second-order object is the posterior-induced distribution over predictive distributions.

Distribution-based formalism: Challenges

- Computing the quantity

$$p(D) = \int p(D \mid \theta) d\pi(\theta)$$

is extremely difficult in practice due to intractability coming from **non-conjugate prior-likelihood** and **high-dimensionality** of θ.

Distribution-based formalism: Challenges

- Computing the quantity

$$p(D) = \int p(D \mid \theta) d\pi(\theta)$$

  is extremely difficult in practice due to intractability coming from **non-conjugate prior-likelihood** and **high-dimensionality** of θ.

- Usually resorts to approximation distribution q_ϕ that maximises the Evidence Lower Bound

$$q_\phi^\star = \arg\max_{q_\phi \in Q} \mathbb{E}_{q_\phi}[-\log p(D \mid \theta)] - \mathrm{D}_{KL}(q_\phi \mid\mid p(\theta))$$

Distribution-based formalism: Challenges

- Computing the quantity

$$p(D) = \int p(D \mid \theta) d\pi(\theta)$$

  is extremely difficult in practice due to intractability coming from **non-conjugate prior-likelihood** and **high-dimensionality** of θ.

- Usually resorts to approximation distribution q_ϕ that maximises the Evidence Lower Bound

$$q_\phi^\star = \arg\max_{q_\phi \in Q} \mathbb{E}_{q_\phi}[-\log p(D \mid \theta)] - \mathrm{D}_{KL}(q_\phi \,||\, p(\theta))$$

- Resembling a distributional empirical risk minimisation + regularisation.

Deep ensembles: a scalable source of predictive distributions

- Train M neural networks **independently**, with initial parameters $\theta^{(1)}, \ldots, \theta^{(M)}$.

Deep ensembles: a scalable source of predictive distributions

- Train M neural networks **independently**, with initial parameters $\theta^{(1)}, \ldots, \theta^{(M)}$.
- For each input x, collect their predictive distributions

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

Deep ensembles: a scalable source of predictive distributions

- Train M neural networks **independently**, with initial parameters $\theta^{(1)}, \ldots, \theta^{(M)}$.
- For each input x, collect their predictive distributions

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

- Average them for the final first-order prediction

$$\bar{p}(\cdot \mid x) = \frac{1}{M} \sum_{m=1}^{M} p_m(\cdot \mid x).$$

Deep ensembles: a scalable source of predictive distributions

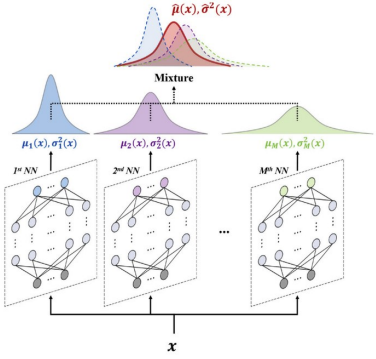
- Train M neural networks **independently**, with initial parameters $\theta^{(1)}, \ldots, \theta^{(M)}$.
- For each input x, collect their predictive distributions

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

- Average them for the final first-order prediction

$$\bar{p}(\cdot \mid x) = \frac{1}{M} \sum_{m=1}^{M} p_m(\cdot \mid x).$$

- Treat disagreement among members as evidence of epistemic uncertainty.



What does ensemble disagreement measure?

Let

$$\bar{p} = \frac{1}{M} \sum_{m=1}^{M} p_m.$$

What does ensemble disagreement measure?

Let

$$\bar{p} = \frac{1}{M} \sum_{m=1}^{M} p_m.$$

The entropy of the averaged prediction decomposes as

$$\underbrace{H(\bar{p})}_{\text{predictive entropy}} = \underbrace{\frac{1}{M} \sum_{m=1}^{M} H(p_m)}_{\text{average member entropy}} + \underbrace{\frac{1}{M} \sum_{m=1}^{M} D_{\mathrm{KL}}(p_m \| \bar{p})}_{\text{ensemble disagreement}}.$$

What does ensemble disagreement measure?

Let

$$\bar{p} = \frac{1}{M} \sum_{m=1}^{M} p_m.$$

The entropy of the averaged prediction decomposes as

$$\underbrace{H(\bar{p})}_{\text{predictive entropy}} = \underbrace{\frac{1}{M} \sum_{m=1}^{M} H(p_m)}_{\text{average member entropy}} + \underbrace{\frac{1}{M} \sum_{m=1}^{M} D_{\mathrm{KL}}(p_m \| \bar{p})}_{\text{ensemble disagreement}}.$$

Within-model uncertainty

Each member may be uncertain about the label.

Between-model uncertainty

Different members may tell different stories about the label.

Distribution-based reading of an ensemble

A finite ensemble is often interpreted as a discrete second-order distribution:

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

Distribution-based reading of an ensemble

A finite ensemble is often interpreted as a discrete second-order distribution:

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

- This treats the ensemble members as sampled predictive distributions.

Distribution-based reading of an ensemble

A finite ensemble is often interpreted as a discrete second-order distribution:

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

- This treats the ensemble members as sampled predictive distributions.
- It enables Bayesian-style uncertainty measures such as **mutual information**, **label-wise variance**, or **distance to a barycentre**.

Distribution-based reading of an ensemble

A finite ensemble is often interpreted as a discrete second-order distribution:

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

- This treats the ensemble members as sampled predictive distributions.
- It enables Bayesian-style uncertainty measures such as **mutual information**, **label-wise variance**, or **distance to a barycentre**.

Distribution-based reading of an ensemble

A finite ensemble is often interpreted as a discrete second-order distribution:

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

- This treats the ensemble members as sampled predictive distributions.
- It enables Bayesian-style uncertainty measures such as **mutual information**, **label-wise variance**, or **distance to a barycentre**.

Takeaway. The distribution-based interpretation asks: how should probability mass be assigned to candidate predictive distributions?

Set-based reading of an ensemble

Instead of treating $\{p_m\}_{m=1}^{M}$ as draws from a hypothetical second-order distribution, we can read it more literally:

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\} \quad \text{is a set of plausible predictive distributions.}$$

Set-based reading of an ensemble

Instead of treating $\{p_m\}_{m=1}^{M}$ as draws from a hypothetical second-order distribution, we can read it more literally:

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\} \quad \text{is a set of plausible predictive distributions.}$$

- The ensemble members are **alternative probabilistic judgements**.

Set-based reading of an ensemble

Instead of treating $\{p_m\}_{m=1}^{M}$ as draws from a hypothetical second-order distribution, we can read it more literally:

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\} \quad \text{is a set of plausible predictive distributions.}$$

- The ensemble members are **alternative probabilistic judgements**.
- The set records which predictions remain **plausible**.

Set-based reading of an ensemble

Instead of treating $\{p_m\}_{m=1}^{M}$ as draws from a hypothetical second-order distribution, we can read it more literally:

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\} \quad \text{is a set of plausible predictive distributions.}$$

- The ensemble members are **alternative probabilistic judgements**.
- The set records which predictions remain **plausible**.
- No **extra commitment** is needed about a probability distribution over ensemble members.

Set-based reading of an ensemble

Instead of treating $\{p_m\}_{m=1}^{M}$ as draws from a hypothetical second-order distribution, we can read it more literally:

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\} \quad \text{is a set of plausible predictive distributions.}$$

- The ensemble members are **alternative probabilistic judgements**.
- The set records which predictions remain **plausible**.
- No **extra commitment** is needed about a probability distribution over ensemble members.

Takeaway. The set-based interpretation asks: which candidate predictive distributions should remain admissible?

Imprecise probability language

- A credal set is a closed and convex set of plausible probability distributions:

$$\mathcal{K} \subseteq \mathcal{P}(\mathcal{Y}).$$

Imprecise probability language

- A credal set is a closed and convex set of plausible probability distributions:

$$\mathcal{K} \subseteq \mathcal{P}(\mathcal{Y}).$$

- It induces lower and upper probabilities for any event $A \subseteq \mathcal{Y}$:

$$\underline{P}(A) = \inf_{p \in \mathcal{K}} p(A), \qquad \overline{P}(A) = \sup_{p \in \mathcal{K}} p(A).$$

Imprecise probability language

- A credal set is a closed and convex set of plausible probability distributions:

$$\mathcal{K} \subseteq \mathcal{P}(\mathcal{Y}).$$

- It induces lower and upper probabilities for any event $A \subseteq \mathcal{Y}$:

$$\underline{P}(A) = \inf_{p \in \mathcal{K}} p(A), \qquad \overline{P}(A) = \sup_{p \in \mathcal{K}} p(A).$$

- This is useful when the available information tells us which probabilities remain plausible, but not how to assign a precise second-order distribution over them.

Partial knowledge can be represented as a set rather than as a single probability.

From ensemble predictions to a credal set

For a fixed input x, the ensemble gives

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

From ensemble predictions to a credal set

For a fixed input x, the ensemble gives

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

For each class k, define bounds

$$\underline{p}_k(x) = \min_m p_{m,k}(x), \qquad \overline{p}_k(x) = \max_m p_{m,k}(x).$$

From ensemble predictions to a credal set

For a fixed input x, the ensemble gives

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

For each class k, define bounds

$$\underline{p}_k(x) = \min_m p_{m,k}(x), \qquad \overline{p}_k(x) = \max_m p_{m,k}(x).$$

Then form the probability-interval credal set

$$\mathcal{K}_x = \left\{ p \in \Delta^{K-1} : \underline{p}_k(x) \le p_k \le \overline{p}_k(x),\ k = 1, \ldots, K \right\}.$$

From ensemble predictions to a credal set

For a fixed input x, the ensemble gives

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

For each class k, define bounds

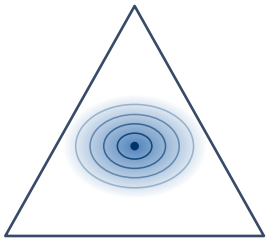
$$\underline{p}_k(x) = \min_m p_{m,k}(x), \qquad \overline{p}_k(x) = \max_m p_{m,k}(x).$$

Then form the probability-interval credal set

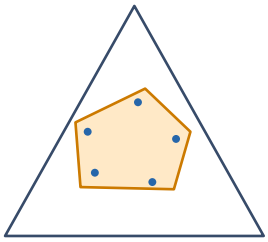
$$\mathcal{K}_x = \left\{ p \in \Delta^{K-1} : \underline{p}_k(x) \le p_k \le \overline{p}_k(x),\ k = 1, \ldots, K \right\}.$$

Takeaway. The same ensemble can support a distribution-based representation $\mathbb{B}_x$ or a credal representation $\mathcal{K}_x$.

Distribution over probabilities vs set of probabilities



or



distribution-based

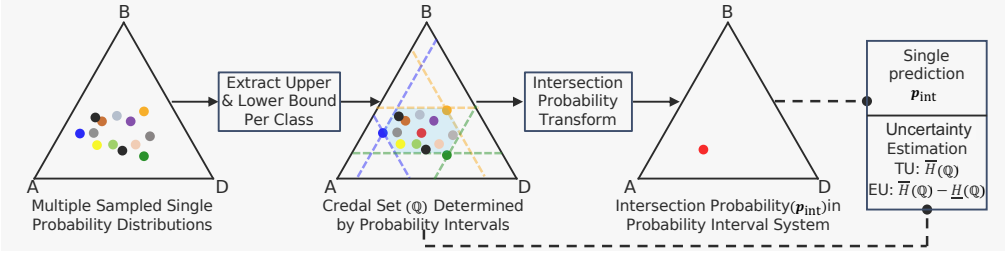
$$\mathbb{B}_x = \frac{1}{M} \sum_m \delta_{p_m}$$

set-based

$$\mathcal{K}_x \subseteq \mathcal{P}(\mathcal{Y})$$

Takeaway. The same raw ensemble predictions can be interpreted either as a discrete second-order distribution or as a set of plausible probabilities.

Credal wrappers: empirical evidence for the set-based reading [Wang et al., 2025]



- A credal wrapper converts any **finite collection** of sampled predictive distributions into a **probability-interval credal set**.
- This makes the set-based interpretation available without changing the underlying predictor.

What the credal wrapper suggests empirically

Table 2: OOD detection AUROC and AUPRC performance (%) of both the classical and credal wrapper version of DEs using EU as the metric. The results are from 15 runs, based on the ResNet-50 backbone. Best scores are in bold.

	CIFAR10 (ID)				CIFAR100 (ID)				ImageNet (ID)	
	SVHN (OOD)		Tiny-ImageNet (OOD)		SVHN (OOD)		Tiny-ImageNet (OOD)		ImageNet-O (OOD)	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Baseline	89.58±0.93	92.29±1.00	86.87±0.20	83.02±0.16	73.83±1.97	84.96±1.25	78.80±0.20	74.68±0.27	65.70±0.41	63.20±0.35
Ours	**93.77±0.60**	**96.06±0.46**	**88.78±0.15**	**86.83±0.23**	**80.22±1.96**	**89.40±1.03**	**81.00±0.16**	**77.16±0.23**	**66.20±0.38**	**63.23±0.34**

Takeaway. There is empirical evidence that treating ensemble disagreement through credal-set concepts can improve uncertainty-centric tasks. But this still begins from the same ensemble predictions.

Section 3

From accidental disagreement to designed disagreement

This part of the talk is based on

Kaizheng Wang, Ghifari Adam Faza, Fabio Cuzzolin, **Siu Lun Chau***, David Moens, and Hans Hallez. "Learning credal ensembles via distributionally robust optimization." International Conference on Machine Learning 2026.*

What most ensemble-to-credal methods still inherit

Common starting point

$$\theta^{(m)} \approx \arg\min_{\theta} \widehat{R}_n(\theta), \qquad m = 1, \ldots, M.$$

All members are trained with the same data and essentially the same objective.

What most ensemble-to-credal methods still inherit

Common starting point

$$\theta^{(m)} \approx \arg\min_{\theta} \widehat{R}_n(\theta), \qquad m = 1, \ldots, M.$$

All members are trained with the same data and essentially the same objective.

Where disagreement comes from

- random initialisation;
- stochastic optimisation;
- different local minima;
- architectural or masking tricks.

What most ensemble-to-credal methods still inherit

Common starting point

$$\theta^{(m)} \approx \arg\min_{\theta} \widehat{R}_n(\theta), \qquad m = 1, \ldots, M.$$

All members are trained with the same data and essentially the same objective.

Where disagreement comes from

- random initialisation;
- stochastic optimisation;
- different local minima;
- architectural or masking tricks.

Takeaway. If all ensemble members solve the same problem, their disagreement may be informative, but it is not deliberately tied to different deployment assumptions.

If we want different stories, ask different questions

Instead of hoping random training gives meaningful diversity, train members to answer different but meaningful objectives.

Standard ensemble diversity

Different seeds, same objective.

$$\widehat{R}_n(\theta)$$

If we want different stories, ask different questions

Instead of hoping random training gives meaningful diversity, train members to answer different but meaningful objectives.

Standard ensemble diversity

Different seeds, same objective.

$$\widehat{R}_n(\theta)$$

Objective-level diversity

Different members encode different robustness assumptions.

$$\widehat{R}_{\delta_m}(\theta), \qquad m = 1, \ldots, M.$$

If we want different stories, ask different questions

Instead of hoping random training gives meaningful diversity, train members to answer different but meaningful objectives.

Standard ensemble diversity

Different seeds, same objective.

$$\widehat{R}_n(\theta)$$

Objective-level diversity

Different members encode different robustness assumptions.

$$\widehat{R}_{\delta_m}(\theta), \qquad m = 1, \ldots, M.$$

Takeaway. The goal is not diversity for its own sake, but disagreement that corresponds to distinct, plausible views of the world.

Distribution shift as the motivating uncertainty

Empirical risk minimisation implicitly assumes

$$P_{\text{test}} = P_{\text{train}}.$$

Distribution shift as the motivating uncertainty

Empirical risk minimisation implicitly assumes

$$P_{\text{test}} = P_{\text{train}}.$$

But in deployment we often only believe

$$P_{\text{test}} \in \mathcal{U}(P_{\text{train}}).$$

Distribution shift as the motivating uncertainty

Empirical risk minimisation implicitly assumes

$$P_{\text{test}} = P_{\text{train}}.$$

But in deployment we often only believe

$$P_{\text{test}} \in \mathcal{U}(P_{\text{train}}).$$

Distributionally robust optimisation replaces average-case training by

$$\min_{\theta} \sup_{Q \in \mathcal{U}(P_{\text{train}})} \mathbb{E}_{(X,Y)\sim Q}\big[\ell(h_\theta(X), Y)\big].$$

Distribution shift as the motivating uncertainty

Empirical risk minimisation implicitly assumes

$$P_{\text{test}} = P_{\text{train}}.$$

But in deployment we often only believe

$$P_{\text{test}} \in \mathcal{U}(P_{\text{train}}).$$

Distributionally robust optimisation replaces average-case training by

$$\min_{\theta} \sup_{Q \in \mathcal{U}(P_{\text{train}})} \mathbb{E}_{(X,Y)\sim Q}\big[\ell(h_\theta(X), Y)\big].$$

Takeaway. DRO makes uncertainty about deployment part of the training objective rather than only part of the post-hoc uncertainty score.

CreDRO: learning a credal ensemble through DRO

- CreDRO trains ensemble members under different robustness levels

$$\delta_1, \ldots, \delta_M.$$

CreDRO: learning a credal ensemble through DRO

- CreDRO trains ensemble members under different robustness levels

$$\delta_1, \ldots, \delta_M.$$

- Member m solves a DRO/CVaR-style objective:

$$\min_{\theta} \max_{w \in \mathcal{W}_{\delta_m}} \frac{1}{N} \sum_{n=1}^{N} w_n \, \ell(h_\theta(x_n), y_n).$$

CreDRO: learning a credal ensemble through DRO

- CreDRO trains ensemble members under different robustness levels

$$\delta_1, \ldots, \delta_M.$$

- Member m solves a DRO/CVaR-style objective:

$$\min_{\theta} \max_{w \in \mathcal{W}_{\delta_m}} \frac{1}{N} \sum_{n=1}^{N} w_n \, \ell(h_\theta(x_n), y_n).$$

- Smaller δ_m gives a more conservative member; $\delta_m = 1$ recovers standard ERM.

CreDRO: learning a credal ensemble through DRO

- CreDRO trains ensemble members under different robustness levels

$$\delta_1, \ldots, \delta_M.$$

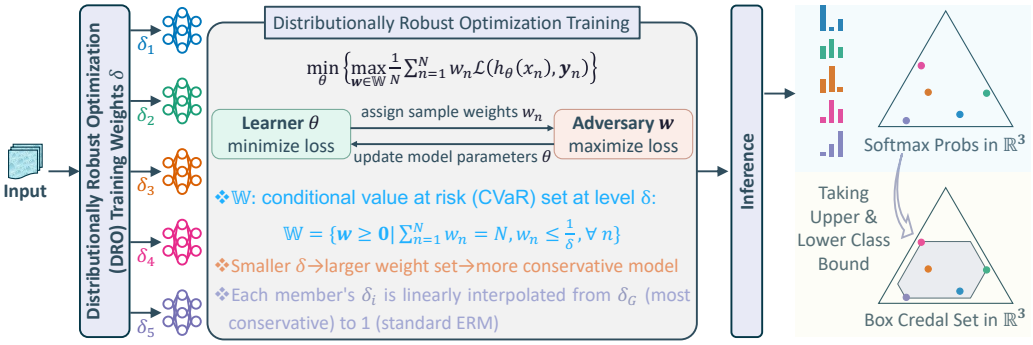
- Member m solves a DRO/CVaR-style objective:

$$\min_{\theta} \max_{w \in \mathcal{W}_{\delta_m}} \frac{1}{N} \sum_{n=1}^{N} w_n \, \ell(h_\theta(x_n), y_n).$$

- Smaller δ_m gives a more conservative member; $\delta_m = 1$ recovers standard ERM.
- The resulting ensemble predictions are converted into a credal set by class-wise bounds.

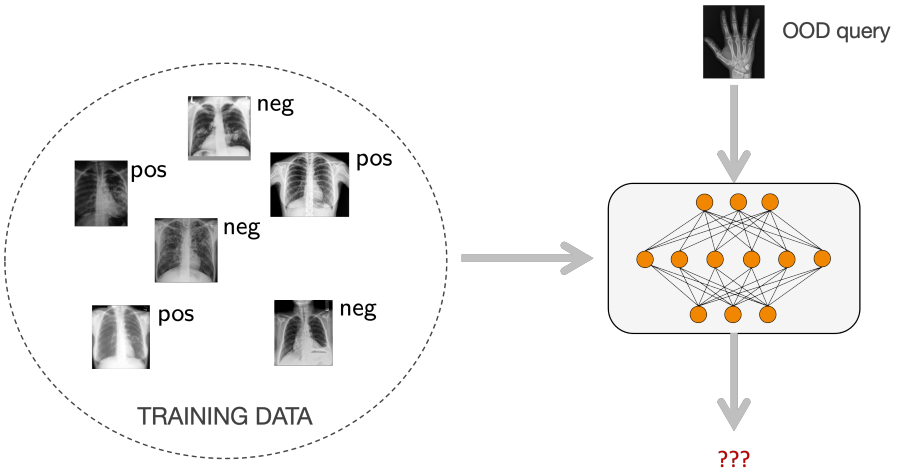
Takeaway. CreDRO makes the ensemble a collection of predictors trained under different plausible distribution-shift assumptions.

CreDRO overview



Wang, Faza, Cuzzolin, **Chau**, Moens and Hallez (ICML 2026 Spotlight).

Evaluation protocol: OOD Detection



OOD Detection

- Perform OOD detection by thresholding the uncertainty:

$$\boxed{x \text{ is OOD} \iff u(x) > \tau,}$$

where τ is chosen using a validation set.

OOD Detection

- Perform OOD detection by thresholding the uncertainty:

$$\boxed{x \text{ is OOD} \iff u(x) > \tau,}$$

  where τ is chosen using a validation set.
- OOD detection is therefore treated as **binary classification problem** using the estimated **epistemic uncertainty**: detect inputs on which the model knows it does not know.

OOD Detection

- Perform OOD detection by thresholding the uncertainty:

$$\boxed{x \text{ is OOD} \iff u(x) > \tau,}$$

  where τ is chosen using a validation set.
- OOD detection is therefore treated as **binary classification problem** using the estimated **epistemic uncertainty**: detect inputs on which the model knows it does not know.
- One particular epistemic uncertainty quantification method is the **entropy imprecision**

$$u_{\mathcal{C}}(x) = \max_{q \in \mathcal{C}(x)} H(q) - \min_{q \in \mathcal{C}(x)} H(q)$$

CreDRO: model disagreement from objective disagreement

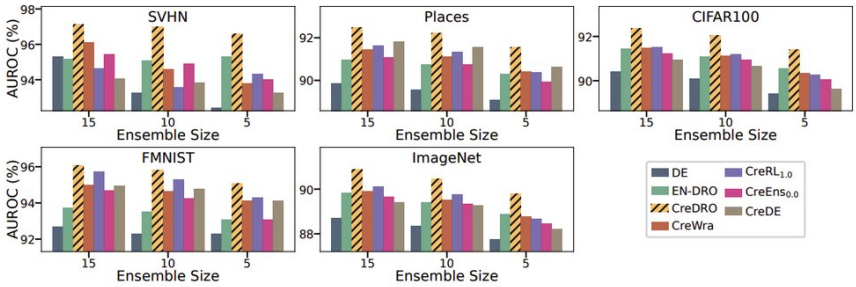


Figure 2. AUROC (%) for OOD detection using EU, across methods and ensemble sizes.

Takeaway. Empirically, disagreement appears more informative when it arises from meaningful differences in training objectives rather than only from random seeds.

Section 4

The elephant in the room: which formalism is "better"?

This part of the talk is based on:

Kaizheng Wang, Yunjia Wang, Fabio Cuzzolin, David Moens, Hans Hallez, and **Siu Lun Chau**. *"Set-based vs Distribution-based Representations of Epistemic Uncertainty: A Comparative Study." The Conference on Uncertainty in Artificial Intelligence 2026.*

The elephant in the room

We have now seen two ways to represent **uncertainty about a predictive distribution**.

Distribution-based

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

A probability distribution over predictive distributions.

Set-based

$$\mathcal{K}_x \subseteq \mathcal{P}(\mathcal{Y}).$$

A set of plausible predictive distributions.

The elephant in the room

We have now seen two ways to represent **uncertainty about a predictive distribution**.

Distribution-based

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

A probability distribution over predictive distributions.

Set-based

$$\mathcal{K}_x \subseteq \mathcal{P}(\mathcal{Y}).$$

A set of plausible predictive distributions.

Which is "better"?

Why this question is hard to answer directly

Suppose we compare a Bayesian neural network with a credal classifier.

Why this question is hard to answer directly

Suppose we compare a Bayesian neural network with a credal classifier.

Many things changed

- learning algorithm;
- architecture;
- optimisation procedure;
- predictive accuracy;
- representation;
- uncertainty measure.

Why this question is hard to answer directly

Suppose we compare a Bayesian neural network with a credal classifier.

Many things changed

- learning algorithm;
- architecture;
- optimisation procedure;
- predictive accuracy;
- representation;
- uncertainty measure.

So what caused the difference?

prediction or **representation** or **measure**?

A comparison is only meaningful if these factors are disentangled.

Why this question is hard to answer directly

Suppose we compare a Bayesian neural network with a credal classifier.

Many things changed

- learning algorithm;
- architecture;
- optimisation procedure;
- predictive accuracy;
- representation;
- uncertainty measure.

So what caused the difference?

prediction or **representation** or **measure**?

A comparison is only meaningful if these factors are disentangled.

Takeaway. To compare second-order representations, we first need a like-for-like construction.

Controlled comparison: same predictions, two representations

Start from exactly the same finite predictive set

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

Controlled comparison: same predictions, two representations

Start from exactly the same finite predictive set

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

Distribution-based representation

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

The ensemble is treated as a discrete second-order distribution.

Controlled comparison: same predictions, two representations

Start from exactly the same finite predictive set

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

Distribution-based representation

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

The ensemble is treated as a discrete second-order distribution.

Credal representation

$$\mathcal{K}_x = \{p \in \Delta^{K-1} : \underline{p}_k(x) \leq p_k \leq \overline{p}_k(x)\}.$$

The same predictions induce class-wise probability bounds.

Controlled comparison: same predictions, two representations

Start from exactly the same finite predictive set

$$\mathcal{P}_x = \{p_1(\cdot \mid x), \ldots, p_M(\cdot \mid x)\}.$$

Distribution-based representation

$$\mathbb{B}_x = \frac{1}{M} \sum_{m=1}^{M} \delta_{p_m(\cdot|x)}.$$

The ensemble is treated as a discrete second-order distribution.

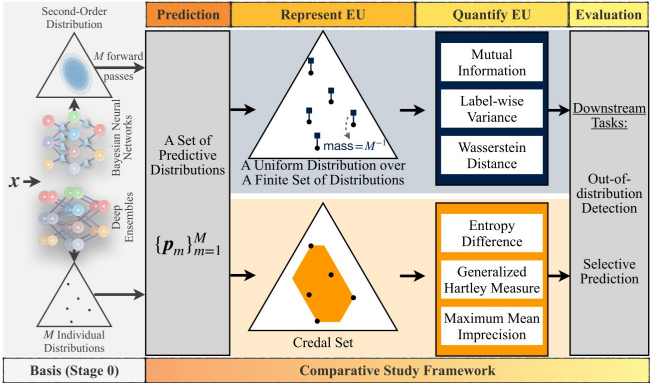
Credal representation

$$\mathcal{K}_x = \{p \in \Delta^{K-1} : \underline{p}_k(x) \leq p_k \leq \overline{p}_k(x)\}.$$

The same predictions induce class-wise probability bounds.

Takeaway. Same network, same predictions, same averaged point prediction. Only the second-order representation changes.

Comparative study framework



Takeaway. The study isolates the representation step before asking how uncertainty scores behave in downstream tasks.

Representation alone is not an uncertainty score

A second-order representation must be queried by a measure.

Distribution-based measures

- Mutual information;
- label-wise variance;
- Wasserstein distance.

Credal measures

- entropy difference;
- generalized Hartley measure;
- **maximum mean imprecision**.

Representation alone is not an uncertainty score

A second-order representation must be queried by a measure.

Distribution-based measures

- Mutual information;
- label-wise variance;
- Wasserstein distance.

Credal measures

- entropy difference;
- generalized Hartley measure;
- **maximum mean imprecision**.

predictive samples $\longrightarrow$ representation $\longrightarrow$ measure $\longrightarrow$ decision.

Representation alone is not an uncertainty score

A second-order representation must be queried by a measure.

Distribution-based measures

- Mutual information;
- label-wise variance;
- Wasserstein distance.

Credal measures

- entropy difference;
- generalized Hartley measure;
- **maximum mean imprecision**.

predictive samples $\longrightarrow$ representation $\longrightarrow$ measure $\longrightarrow$ decision.

Takeaway. "Distribution or set?" is incomplete. The measure used to interrogate the representation matters.

Downstream tasks used for evaluation

Selective prediction

- Rank test samples by estimated epistemic uncertainty.
- Reject the most uncertain samples.
- Good uncertainty should improve accuracy on retained samples.

metric: AUARC

OOD detection

- Treat ID and OOD samples as two classes.
- Use epistemic uncertainty as the detection score.
- Good uncertainty should assign higher scores to OOD samples.

metric: AUROC

Downstream tasks used for evaluation

Selective prediction

- Rank test samples by estimated epistemic uncertainty.
- Reject the most uncertain samples.
- Good uncertainty should improve accuracy on retained samples.

metric: AUARC

OOD detection

- Treat ID and OOD samples as two classes.
- Use epistemic uncertainty as the detection score.
- Good uncertainty should assign higher scores to OOD samples.

metric: AUROC

Takeaway. There is no ground-truth epistemic uncertainty label, so evaluation happens through decisions.

Experimental scope

Predictive models

- stochastic variational inference;
- Monte Carlo dropout;
- deep ensembles;
- BatchEns, MaskEns, PackEns.

Benchmarks

- CIFAR10 with SVHN and FMNIST as OOD;
- Camelyon17 medical image shifts;
- SeaShip and SeaShip corruption/OOD variants;
- reciprocal FMNIST/SVHN experiments in the appendix.

Experimental scope

Predictive models

- stochastic variational inference;
- Monte Carlo dropout;
- deep ensembles;
- BatchEns, MaskEns, PackEns.

Benchmarks

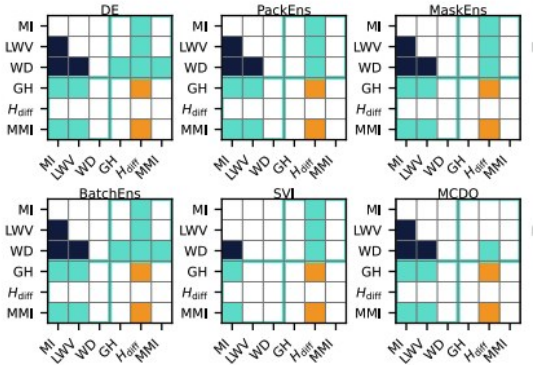
- CIFAR10 with SVHN and FMNIST as OOD;
- Camelyon17 medical image shifts;
- SeaShip and SeaShip corruption/OOD variants;
- reciprocal FMNIST/SVHN experiments in the appendix.

6 predictive model families, 6 uncertainty measures, 14 benchmarks, 10 independent runs per setting.

Selective prediction: differences are visible but compressed

- A shaded cell means the row measure significantly outperforms the column measure.
- Selective prediction often gives very similar rejection sets.
- If base accuracy is already high, performance gaps are naturally compressed.

Takeaway. For selective prediction, Wasserstein distance and credal GH/MMI are strong, but average AUARC differences are small.

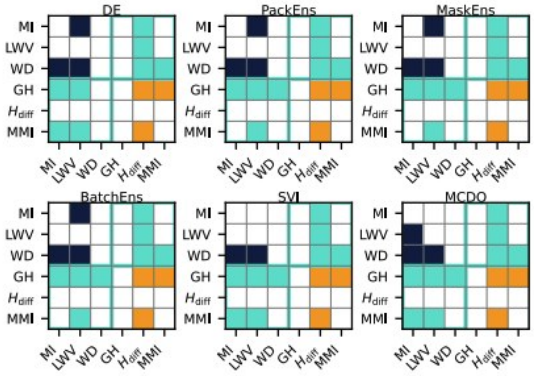


(a) Selective prediction on in-distribution Camelyon17 test data

OOD detection: representation differences become clearer

- OOD detection depends directly on how uncertainty reacts to distribution shift.
- The task is therefore more sensitive to representation and measure choice.
- The generalized Hartley measure performs especially strongly for the credal representation.

Takeaway. OOD detection exposes representational differences more clearly than selective prediction.



(c) OOD detection on SeaShip v.s. SeaShip-O

Finding 1: no representation is uniformly superior

The question "distribution or set?" has no unconditional answer.

- The rankings **changes with the uncertainty measure**.

Finding 1: no representation is uniformly superior

The question "distribution or set?" has no unconditional answer.

- The rankings **changes with the uncertainty measure**.
- It also changes with **dataset, predictive model, and downstream task**.

Finding 1: no representation is uniformly superior

The question "distribution or set?" has no unconditional answer.

- The rankings **changes with the uncertainty measure**.
- It also changes with **dataset, predictive model, and downstream task**.
- Therefore, claims about second-order representations should always state the full evaluation pipeline.

Finding 1: no representation is uniformly superior

The question "distribution or set?" has no unconditional answer.

- The rankings **changes with the uncertainty measure**.
- It also changes with **dataset, predictive model, and downstream task**.
- Therefore, claims about second-order representations should always state the full evaluation pipeline.

$$\text{representation} + \text{measure} + \text{task} + \text{data}.$$

Finding 2: OOD detection is the sharper probe

Selective prediction

- rejects uncertain samples;
- different scores may reject similar samples;
- high baseline accuracy compresses AUARC gains.

OOD detection

- directly tests response to distribution shift;
- uncertainty score is the classifier score;
- differences in representation become more visible.

Finding 2: OOD detection is the sharper probe

Selective prediction

- rejects uncertain samples;
- different scores may reject similar samples;
- high baseline accuracy compresses AUARC gains.

OOD detection

- directly tests response to distribution shift;
- uncertainty score is the classifier score;
- differences in representation become more visible.

Takeaway. New uncertainty representations should be evaluated across multiple downstream tasks, not only on one convenient metric.

Finding 3: metric choice matters as much as representation

Distribution-based side

- Wasserstein distance is consistently strong.
- It captures geometric dispersion of predictive distributions in the simplex.
- MI and label-wise variance can miss aspects of this geometry.

Finding 3: metric choice matters as much as representation

Distribution-based side

- Wasserstein distance is consistently strong.
- It captures geometric dispersion of predictive distributions in the simplex.
- MI and label-wise variance can miss aspects of this geometry.

Credal side

- Generalized Hartley is consistently strong.
- It reflects the structure of the set-valued prediction.
- Entropy difference is popular but often performs poorly.

Finding 3: metric choice matters as much as representation

Distribution-based side

- Wasserstein distance is consistently strong.
- It captures geometric dispersion of predictive distributions in the simplex.
- MI and label-wise variance can miss aspects of this geometry.

Credal side

- Generalized Hartley is consistently strong.
- It reflects the structure of the set-valued prediction.
- Entropy difference is popular but often performs poorly.

Takeaway. Better representations still need better uncertainty measures; the two cannot be separated empirically.

What this suggests for future work

1. **Do not stop at calibration.** Calibration is useful but it is not a **representation of epistemic uncertainty**.

What this suggests for future work

1. **Do not stop at calibration.** Calibration is useful but it is not a **representation of epistemic uncertainty**.
2. **Make the second-order object explicit.** Is it a distribution over predictions, a set of predictions, or something else?

What this suggests for future work

1. **Do not stop at calibration.** Calibration is useful but it is not a **representation of epistemic uncertainty**.
2. **Make the second-order object explicit.** Is it a distribution over predictions, a set of predictions, or something else?
3. **Design disagreement when possible.** Ensembles should ideally disagree because they **encode different meaningful assumptions**.

What this suggests for future work

1. **Do not stop at calibration.** Calibration is useful but it is not a **representation of epistemic uncertainty**.
2. **Make the second-order object explicit.** Is it a distribution over predictions, a set of predictions, or something else?
3. **Design disagreement when possible.** Ensembles should ideally disagree because they **encode different meaningful assumptions**.
4. **Report the full uncertainty pipeline.** Representation, measure, task, dataset, and model all matter.

Takeaway. Progress in uncertainty-aware learning is driven by both representational advances and principled uncertainty measures.

Final takeaway

One probability is not enough.

First-order prediction

What is the probability of the outcome?

Second-order representation

How uncertain are we about that probability?

Decision

How should uncertainty change what we do?

Takeaway. The right object is not just a confidence score, but a representation that guides uncertainty-aware decisions.

Section 5

Backup slides

Backup: distribution-based uncertainty measures

Given $\mathbb{B}_x = \frac{1}{M}\sum_m \delta_{p_m}$ and $\bar{p} = \frac{1}{M}\sum_m p_m$:

$$\mathsf{MI}(\mathbb{B}_x) = H(\bar{p}) - \frac{1}{M}\sum_{m=1}^{M} H(p_m),$$

$$\mathsf{LWV}(\mathbb{B}_x) = \sum_{k=1}^{K} \bar{p}_k(1-\bar{p}_k) - \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K} p_{m,k}(1-p_{m,k}),$$

$$\mathsf{WD}(\mathbb{B}_x) = \min_{p^\star \in \Delta^{K-1}} \frac{1}{2}\sum_{m=1}^{M} \|p_m - p^\star\|_1.$$

Backup: credal uncertainty measures

Given $\mathcal{K}_x \subseteq \Delta^{K-1}$:

$$\mathsf{Hdiff}(\mathcal{K}_x) = \max_{p \in \mathcal{K}_x} H(p) - \min_{p \in \mathcal{K}_x} H(p),$$

$$\mathsf{GH}(\mathcal{K}_x) = \sum_{A \subseteq \mathcal{Y}} m_{\mathcal{K}}(A) \log_2 |A|,$$

$$\mathsf{MMI}(\mathcal{K}_x) = \sup_{A \subseteq \mathcal{Y}} \left\{ \overline{P}(A) - \underline{P}(A) \right\}.$$

Takeaway. In binary classification, MMI reduces to the interval length. GH is expressive, but scales over subsets of the label space.

Backup: numerical summary of the main paper

Selective prediction

	Camelyon	Shift	SeaShip	CIFAR
MI	.960	.964	.989	.981
LWV	.963	.967	.990	.982
WD	**.963**	**.968**	**.990**	**.982**
GH	.963	.968	.990	.982
Hdiff	.959	.961	.985	.976
MMI	.963	.968	.990	.982

OOD detection

	Ship-O	Ship-C	SVHN	FMNIST
MI	.869	.825	.766	.864
LWV	.866	.830	.769	.847
WD	.882	.851	.817	.873
GH	**.886**	**.856**	**.830**	**.880**
Hdiff	.852	.793	.780	.874
MMI	.877	.846	.804	.863

Numbers are averages across six predictive model families and ten runs per configuration.

Backup: what the comparison deliberately does not claim

- It does not claim that one representation universally dominates the other.
- It does not evaluate every possible second-order formalism: Dirichlet-based and random-set models are not the focus.
- It does not replace theoretical analysis of semantics.
- It isolates a specific empirical question: what changes when the same finite predictive set is represented differently?

Takeaway. The contribution is a controlled comparison, not a final verdict on all forms of epistemic uncertainty.

Selected references

- Lakshminarayanan, Pritzel and Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. NeurIPS 2017.
- Hullermeier and Waegeman. Aleatoric and epistemic uncertainty in machine learning. Machine Learning 2021.
- Wang et al. Credal wrapper of model averaging for uncertainty estimation in classification. ICLR 2025.
- Sale, Bengs, Caprio and Hullermeier. Second-order uncertainty quantification: a distance-based approach. ICML 2024.
- Chau, Caprio and Muandet. Integral imprecise probability metrics. NeurIPS 2025.
- Wang et al. Learning credal ensembles via distributionally robust optimization. ICML 2026.
- Wang et al. Set-based vs. distribution-based representations of epistemic uncertainty: a comparative study. 2026.

References I

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal wrapper of model averaging for uncertainty estimation in classification. In *The Thirteenth International Conference on Learning Representations*, 2025.